DOCUMENT RESUME

ED 382 650 TM 023 084

AUTHOR	Longford, Nicholas T.
TITLE	Models for Scoring Missing Responses to Multiple-Choice Items. Program Statistics Research Technical Report No. 94-1.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-94-9
PUB DATE	Mar 94
NOTE	26p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Algorithms; Classification; Coding; *Models; Motivation; *Multiple Choice Tests; *Scoring; *Test Construction; Test Items
IDENTIFIERS	*Missing Data; *National Assessment of Educational Progress

ABSTRACT

This study is a critical evaluation of the roles for coding and scoring of missing responses to multiple-choice items in educational tests. The focus is on tests in which the test-takers have little or no motivation; in such tests omitting and not reaching (as classified by the currently adopted operational rules) is quite frequent. Data from the 1991 National Assessment of Educational Progress (NAEP) Reading Assessment of 17-year-olds are used in analyses and illustrative examples. Alternative rules for scoring based on hypothesized behavior of the test-takers are proposed. The approach for incorporation of information about missing responses relies on a model relating knowledge categories (know or does not know) to the response categories (correct, incorrect, omitted, not reached, multiple). A computational algorithm is described that requires no new technology to be developed. Two tables and two figures describe the scoring approach. (Contains 2 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 382 650

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. COLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

RR-94-9

Models for Scoring Missing Responses to Multiple-Choice Items

Nicholas T. Longford
Educational Testing Service

PROGRAM
STATISTICS
RESEARCH

Technical Report No. 94-1

Educational Testing Service
Princeton, New Jersey 08541

TM023084

BEST COPY AVAILABLE

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Models for Scoring Missing Responses to Multiple-Choice Items

Nicholas T. Longford
Educational Testing Service

Program Statistics Research
Technical Report No. 94-1

Research Report No. 94-9

Educational Testing Service
Princeton, NewJersey 08541

March 1994

Copyright © 1993 by Educational Testing Service. All rights reserved.

Models for scoring missing responses to multiple-choice items

N.T. Longford

Abstract

This is a critical evaluation of the rules for coding and scoring of missing responses to multiple-choice items in educational tests. The focus is on tests in which the test-takers have little or no motivation; in such tests omitting and not reaching (as classified by the currently adopted operational rules) is quite frequent. Data from the 1991 NAEP Reading Assessment of 17-year-olds are used in analyses and illustrative examples. Alternative rules for scoring based on hypothesized behaviour of the test-takers are proposed.

Some key words: EM algorithm; item-response theory; missing data; mixture models.

Introduction

No amount of effort and ingenuity expanded on instructing and motivating test-takers can ensure that they dilligently attend to each item and respond accurately to background and experiential items, and respond to cognitive items to the best of their ability. This problem is particularly prevalent when the test-takers have little or no stake in the outcome of the test administration. Large-scale educational surveys, such as NAEP, NALS, and the like, are cases in point.

While accuracy of the responses to background questionnaire items can to some extent be improved by cross-referencing of the items, information about *knowledge* of the correct response to an item is very fractured. If a test-taker fails to respond to an item he/she might or might not have responded correctly had his/her acumen been applied to the item fully. The pattern of responses may provide some clues about this hypothetical response. For instance, if the neighbouring items (the previous and the next ones) were responded to, it is very likely that the item in question was attended by the test-taker, but he/she did not know the response. If a sequence of consecutive items is not responded to, it is likely that the items, with the possible exception of a few items at the beginning of the sequence, were skipped (not attended/inspected). On the other hand, non-missing responses with a regular pattern (e.g., the first response option marked for each item) imply that the test-taker either did not inspect the test items (except possibly a few items at the beginning of the test), or did not apply his/her acumen in responding to them. Of course, such patterns can be detected, at least in principle, by direct inspection of the responses.

This paper offers a critical evaluation of the rules for coding and scoring of missing responses to multiple-choice items in the National Assessment of Educational Progress (NAEP). Although a number of issues are discussed in the specific context of NAEP, most of them are applicable to other large-scale educational tests in which the test-takers have little at stake.

The next section describes the operational rules and conventions for coding and scoring of responses to cognitive items in NAEP, and outlines the steps in processing of these scores. The following section considers alternative schemes for incorporating missing responses in item-response models using the EM algorithmic approach. Then some of the mental processes that may lead to missing responses are discussed.

Operational rules for classification and scoring

A typical multiple-choice item consists of a reading passage (instruction) concluded by a question, and a small number of response options, one of which is correct and the others are incorrect. The items are organized in blocks, each block administered under standardized conditions common to all test-takers. The conditions usually include administration time and availability of tools, such as calculators or dictionaries. For instance, test-takers may have 40 minutes to respond to a block of twelve items. A test may consist of several blocks, each timed separately. In the pencil-and-paper mode of administration there is a natural ordering of the items. An integral component of the item (block, or test) is the instruction or suggestion how to respond when the correct response is not known. For example, the test-takers may be instructed not to respond at all or to select a response arbitrarily when they do not know which response is correct.

In NAEP, as in most other educational tests, the response, or its absence, to a cognitive multiple-choice item is classified as

- correct (C)
- incorrect (N)
- omitted (O)
- not reached (NR)
- multiple response (M).

The response to an item is classified as *omitted* if it is missing, and at least one of the later items in the block has a presented response (correct, incorrect, or multiple). Each response from a contiguous sequence of missing responses to items which contain the last item of the block is classified as *not reached*. Multiple responses to an item can be effectively dealt with by randomly selecting one of the marked responses, assuming that the test-taker narrowed down the range of possible correct answers to these options. The selected response is then classified as either correct or incorrect. When each marked response is incorrect the selection is irrelevant.

A correct response is scored one point, and an incorrect one no points. An 'omitted' response to a multiple-choice item with K response options is scored $1/K$. A 'not-reached' response is regarded as *non-informatively* missing, given the pattern of responses up to the first item of the 'not-reached' sequence. Without delving into technical details, this can be interpreted as follows: Suppose test-taker T has the sequence of 'not reached' responses to items $I+1$, $I+2$, ..., M. Then the test-taker's score for a 'not reached' item J ($I \leq J \leq M$) is, in essence, equal to the mean of the scores for item J of all the test-takers who have the same pattern of responses to items 1, 2, ..., I as test-taker T but who 'reached' item J.

An objective of the test administration is to make inferences about the proportions of the test-takers who *know* the correct answers to the items, about summaries of other variables for such test-takers, and to make inferences about *proficiency scores* underlying the subject area represented in the block. The proficiency scores are linked to the probabilities of knowing the correct responses. The class of methods for estimation of proficiency scores is known as *item-response theory*.

We classify the test-taker's knowledge of an item as

- knows the correct response (K)
- does not know (DK).

Table 1 describes the relationship of the response categories to the knowledge categories under some simple but realistic assumptions. The response of a test-taker who does not know the correct response may fall into either of the four categories. A test-taker who knows the correct response either does select the correct response or abandons the block before reaching the item. This model has two important aspects. First, no intermediate states between 'knowing' and 'not-knowing' are considered. Second, the knowledge state is assumed not to depend on extraneous factors, such as the time elapsed while the item is attended, experience with the previous item, and the like.

It would be useful to classify 'not-reached' responses according to whether the item was attended or the block was abandoned prior to the item. The former would be more appropriate to classify as an 'omit'. Thus, a sequence of 'not-reached' items may start with one or several (missing) responses even though the items were inspected.

The current classification of responses relies on the assumption that the test-takers attend to the items in a *linear* fashion, that is, in the order of presentation. The scoring rule is based on the assumption that the 'omitted' response to an item was given after inspection of the item, while a 'not reached' response was given without inspecting the item. This assumption is clearly fallacious. For example, the response to the last item can never be classified as 'omitted', even though it is feasible that a test-taker inspects the last item and fails to respond to it.

In large-scale administrations of educational tests used primarily for survey purposes, such as NAEP, proficiency scores for individual test-takers are of lesser importance than *mean proficiency scores* for subpopulations. In such settings proficiency scores for test-takers are at best intermediate steps in estimating (sub-)population summaries, such as means. An alternative strategy would be to bypass this intermediate step and estimate the (sub-)population summaries directly, taking into account the uncertainty due to non-response. This would make the issue of imputing responses for individual test-takers moot. Such non-

response could be integrated in a framework of inference in which the fact that only a fraction of the entire population is tested is regarded as another source of uncertainty.

The next section summarizes item-response models for tests with multiple-choice items. This is followed by an outline of a formal probabilistic mixture model for missing responses, distinguishing between types of missingness. Then an EM algorithm for estimation with such a model is described.

Item-response models

Item-response models describe the relationship of the probability of knowing the correct response to an item to the *proficiency* of the test-taker and the properties of the item. Typical item-response models assume that each of a sample of J test-takers responds to each of a set of I multiple-choice items. There is no formal provision for missing responses. In the implementation for NAEP, scores are imputed for the missing responses as indicated in the previous section. A particular challenge addressed by the item response models is the diffused character of the data. By its nature, each item is an imperfect representation of the tested domain (Critical Thinking, Arithmetic, or the like), the score for a presented response to each item is very coarse (zero or unity), and even this score is contaminated; a test-taker may guess the correct answer by chance, with a non-trivial probability.

This line of thought naturally leads to the discussion of incomplete information and the EM algorithm. In the EM algorithmic framework a hypothetical *complete data* set is considered, and the dataset available, the *incomplete data* set, is formed from the complete one by discarding part of the data. It is advantageous to select the complete data set so that if it were available, its analysis would be relatively simple. Estimation of the model parameters using the incomplete data is then accomplished by iterations of the EM algorithm. Each iteration comprises two steps. The E-step involves, in essence, estimation of the *missing data* which complement the incomplete data, and the M-step is

the algorithm applicable for the complete data, in which the missing data are replaced by their 'estimates.'

An EM algorithm, as applicable to 'omitted' and 'not reached', would involve 'estimation' of these missing responses (i.e., evaluating the conditional probabilities of the correct responses), and these conditional probabilities, given the data and current estimates of the model parameters, would be used in place of the missing data. This is not a feasible proposition, though, because fitting an item-response model (even with complete data) is computationally rather demanding. In the EM algorithm such a model would have to be fitted once in every iteration. Also, missing information appears to be handled rather inefficiently in this algorithm. An 'omitted' response is usually assumed to correspond to not knowing the correct response with certainty, yet this is not reflected in the complete data set of the EM algorithm.

In the M-step a function of a missing data point is replaced by the conditional expectation of that function (as opposed to the function of the conditional expectation). For instance, let p be the conditional expectation imputed for a missing response. Then the square of this missing data point would be replaced in the M-step by the conditional expectation of the square, $(1-p)\cdot 0^2+p\cdot 1^2 = p$, not by p^2.

As an alternative, consider the complete data set comprising of the binary outcomes K (knows the correct response) and DK (does not know). Now the M-step consists of a relatively easy task of fitting a two-parameter item-response model, and estimating the (population) probabilities of *guessing*, i.e., the conditional probabilities of correct response given that the correct response is not known. The E-step consists of evaluating the conditional probabilities of knowing given the (presented or missing) response, and the current estimates of all the parameters.

Such an E-step requires a model for missingness that relates the probabilities of knowing to the missing responses. If the present classification is taken at face value, a model is required only for 'not-reached' responses.

Missing response mechanisms

Several reasons for not responding can be readily identified. First, a test-taker may be uncertain about which response to mark, and ends up not marking either of them. An extreme version of this behaviour is when a test-taker inspects each item cursorily, and then responds only to the items for which he/she is confident of knowing the correct responses. Of course, the test-taker's judgement about his/her ability to identify the correct response may not be accurate. Next, a test-taker may be discouraged from continuing by a sequence of items for which he/she does not know the correct response. Also, a very bright test-taker may be discouraged by too easy items. Thus, it is feasible that a test-taker who did not attend to an item does know the correct response, and would have responded correctly, had he/she attended to it. Conceivably, a test-taker may know the correct response to several items, but identifying the responses requires such effort that he/she is not prepared to attend to them fully.

In practice, the cause of a missing response is not known. Partial information is available when linear order of attendance to the items is assumed. For instance, if a contiguous sequence of items is not responded to, it is likely not to have been attended to. If a response is missing but the responses to the preceding and following items are not, the missing response is likely to be the consequence of not knowing. However, in a sequence of missing responses to consecutive items, the first few responses may be due to not knowing, and the rest due to not attending. The point at which the block was abandoned is not known.

Clearly, information about attendance of items would enhance information about knowledge of the items, and would promote a more appropriate classification of responses. In other words, it would reduce the amount of missing information. This could be arranged, for example, by asking the test-takers to mark a box after inspecting the item. Of course, test-takers might mark such a box without inspecting the item. Also, instructions about such boxes may

confuse some test-takers, or distract them from the principal goal of the test. In computer-based administration the issue of attending items is moot because the time elapsed while an item is displayed on the screen can be recorded. Nevertheless, an item may be displayed for a long time, without being attended to (for example, the last item 'attended'). The order in which the items are presented can also be recorded with the computer administration.

Test-takers could be instructed to use two kinds of symbols: to circle the response options they believe are correct (one per item), and to cross out those they believe are incorrect. Now, if the correct response, and no other, is circled, it is scored one point; if another response is circled, it is scored no points; if a correct and some incorrect responses are circled, either no points are scored, or one of the responses is selected at random. Information from the crossed-out responses would be used only when the correct response is not circled; the test-taker may have narrowed down the choice of the correct response, or he/she crossed out the correct response. The score for either response pattern can be defined in a natural way. This approach opens up possibilities for a variety of partial-credit scoring schemes.

Models for missing responses

Let ϕ_i be the probability of abandoning the block of items immediately after item $i-1$ (at the beginning of the block, for $i=1$). That is, the response to item $i-1$ is C or N, unless $i=1$, and the responses to items $i+1, \ldots, I$, are all missing. Let $f_1 = \phi_1$ and, recursively,

$$f_i = f_{i-1} + (1 - f_{i-1})\phi_{i-1};$$

f_i is the (cumulative) probability of abandoning the test at any point prior to item i. Further, let r_i be the conditional probability of being able to respond to item i correctly, given that the item is not reached. Suppose the decision to abandon the block is not influenced by the abandoned section of the block and that the conditional probability of knowing the correct response, given not

reaching, does not depend on the item at which the block was abandoned. In some contexts it is reasonable to assume that $r_i < p_i$, where p_i is the probability of knowing the correct response. That is, those who do not reach have lower probabilities of knowing the correct response than those who reach the item. Let p_i' be the conditional probability of knowing the correct response given that the item is reached;

$$p_i' = \frac{p_i - f_i r_i}{1 - f_i} .$$

Further, let o_i be the conditional probability of omitting, given that the item is inspected (reached) and the correct response is not known. Thus, the (unconditional) probability of omitting is

$$(1 - f_i)(1 - p_i')o_i .$$

Next, let c_i be the conditional probability of guessing the correct response given that the item is reached, the correct response is not known, and the item is not omitted. Then the probability of the correct response is

$$P_i(C) = (1 - f_i)\{p_i' + (1 - p_i')(1 - o_i)c_i\},$$

the probability of an incorrect response is

$$P_i(N) = (1 - f_i)(1 - p_i')(1 - o_i)(1 - c_i),$$

and the probability of omitting is

$$P_i(O) = (1 - f_i)(1 - p_i')o_i .$$

Note that $P_i(C) + P_i(N) + P_i(O) + f_i = 1$.

Identification and estimation

Clearly, the probability r_i cannot be identified. The remaining four parameters associated with item i, f_i, p_i, o_i, and c_i, cannot be identified from the counts of correct, incorrect, omitted, and not reached responses, because the counts have

a fixed total. Test-takers who did not reach item i have abandoned the block earlier (at item $i' < i$), omitted item i and all consequent items in the block, or omitted a contiguous sequence of items containing item i, and then abandoned the block.

There are no omissions at the last item of the block because all no-responses are classified as not reached. Also, no-responses near the end of the block are more likely to be classified as not reached because the chances of inspecting and not knowing each item at the end of the block increase.

It is instructive to consider two subsets of the responses to item i:

- test-takers who responded to the last item;
- test-takers who failed to reach the last item but reached item i.

In the Reading blocks the latter subset tends to omit more and have lower proportions of correct responses to most items, even after conditioning out the omitted responses. Figure 1 contains the plots of proportions correct, incorrect, and omitted, for the two subsets and each item. The solid, dotted, and dashed lines join the respective probabilities of correct, incorrect, and omitted responses for the items. The thick lines represent the subset of test-takers who reached the last item, the thin line those who reached the item i, but not the last item. The diagram supports the hypothesis that the present classification of 'not reached' represents a mixture of two distinct behaviours:

- inspected and did not respond;
- abandoned prior to reaching the item.

There are several ways of ensuring identification of the probabilities associated with item i. For instance, if the conditional probability of omit given response to the previous item and no response to any of the following items (i.e., not reached starts here) were known probabilities of omission would be estimable.

Conditional independence of the responses, given item and test-taker is a key assumption in all standard item-response models. Not reaching represents an important violation of this assumption. Also, it is conceivable that a response is more likely to be omitted after an omission, and that not reaching is more frequent among those who tend to omit more. These hypotheses can be informally tested by suitable summaries, such as those in Figure 1. Another important departure arises in tests in which speededness is an issue. Test-takers who work slowly and dilligently may respond well to the earlier items, but run out of time and not reach the later items. Conceivably, had they had plenty of time, they would have responded even to the later items better than test-takers with the same responses on the earlier items who did reach the end of the block.

In Figure 2 the relative proportions of the test-takers who failed to reach the end of the block are plotted for each block. In the left-hand plot the proportion scale and in the right-hand plot the logit scale are used. Notable are the steep drops at the end of most blocks. This implies, as conjectured by Wainer (1985), that among those test-takers who are classified as 'not reached the last item' a large fraction did attend the last item.

Models for omission

It is reasonable to assume that test-takers have varying 'propensity' to omit an item. Also this propensity may be associated with proficiency; less able test-takers are more likely to omit. It is important to distinguish two kinds of probabilities related to omission. The absolute (unconditional) probability of omission is bounded by zero and the probability of not knowing the correct response. The conditional probability of omitting, given 'do not know', can attain any value in the range [0,1]; for instance, a test-taker may omit every item for which he/she does not know the correct response.

The frequencies of omission vary a great deal from item to item. There is a perceptible trend toward higher frequencies for later items, with the obvious exception of the last (and sometimes also the penultimate) item.

Insight into patterns of omission can be gained by summarizing the pattern for each test-taker, and by crosstabulating the type of response (presented response/omit/not-reached). Table 2 contains an example. There is a positive association of omitting consecutive items. Those who omit an item are more likely to omit the next item also. Note that the table contains 'structural' zeros as in the row 'not-reached'. If a test taker does not reach an item he/she fails to reach the next item also. Also, omission cannot be immediately followed by 'not reached'.

The number of omits is very small for most items. Table 2 is unusual in that it has relatively large entries in each cell (other than structural zeros). The number of omits at item 6, almost 40 per cent, is extemely high. Other blocks contain at most one item each with omission rate higher than ten per cent. The omission rates for most items are less than one per cent. Nevertheless, positive association of omission behaviour is transparent; those who omit an item are more likely to omit the next item also. This hypothesis can be formally tested by evaluating the log-odd ratios for all the contingency tables (such as in Table 2) which have sufficiently large entries, by pooling these tables (or all the tables), or by application of the Mantel-Haenszel method for estimating the common log-odds ratio. For example, the log-odds ratios for the 2×2 contingency tables for pairs of items with more than 10 omits each are 0.92 (items 1 and 2), 1.09 (5 and 6), and 1.67 (9 and 10) in block C (comprising 11 items); 1.87 (3 and 4), 2.91 (4 and 5), −0.09 (6 and 7), and 1.52 (7 and 8) in block D (9 items); 3.13 (4 and 5), 1.79 (5 and 6), 2.57 (7 and 8), 3.11 (8 and 9) 1.65 (9 and 10), and 1.12 (10 and 11) in block E (12 items), and so on. Note that the log-odds ratios are large even for the penultimate pair (the log-odds ratio for the last pair is always equal to zero because responses to the last item are never classified as 'omitted').

This provides clear evidence that test-takers differ in their rates of omission. Similarly, it can be observed that lower ability test-takers and those who do not reach the last item ('abandon' the block) tend to omit more frequently.

Naive estimates of these rates are tainted by the imperfect classification of item responses.

Alternative IRT models

In this section alternative treatments of missing responses are proposed. The alternatives aim at improved classification of missing responses.

EM algorithm with the three-parameter IRT

We adopt the working assumption that had each test-taker responded to every item the three-parameter IRT model would be appropriate. Thus the probability of a correct response is modelled by the equation

$$\mathrm{P}(R_{ti} = \mathrm{C}) \;=\; c_i + (1 - c_i)\mathrm{logit}^{-1}(a_i + b_i\theta_t), \tag{1}$$

where $\mathrm{logit}(p) = \log\{p/(1-p)\}$ for $p \in (0,1)$, so that $\mathrm{logit}(x)^{-1} = \exp(x)/\{1+\exp(x)\}$ for $x \in (-\infty, +\infty)$. The *proficiencies* θ_t for test-takers $t = 1, \ldots, T$ are either assumed to be unknown constants, or a random sample from a given distribution (usually the standard normal). The item parameter vectors (a_i, b_i, c_i), $i = 1, \ldots, I$, are either assumed to be unknown vectors of constants, or are a random sample from a trivariate (usually normal) distribution with unknown mean vector and variance matrix. The quantities (a_i, b_i, c_i) are interpreted as difficulty (related to the probability of 'K' for the average test-taker), discrimination (related to the increment of the probability of 'K' for an infinitesimal unit of proficiency), and the (conditional) probability of guessing (given 'DK'), respectively.

Missing data, that is, 'omits' and 'not-reached', create an obvious problem, primarily because missingness is associated with the knowledge state, and this association is not fully specified. For instance, an 'omitted' response implies 'DK' with certainty, but a 'not-reached' does not. However, a 'not reached' response may in fact be an omitted response.

This problem can be formulated within the EM algorithmic framework, see Dempster, Laird, and Rubin (1977). A hypothetical *complete* dataset is considered, in which each test-taker responds to each item (and each response is classified as either 'C' or 'N'). The observed (available) data is referred to as the *incomplete* dataset. *Missing* data, the complement of the incomplete data, can be thought of as the hypothetical responses of the test-takers to the items that they failed to respond (had they responded their responses would have been ...).

The EM algorithm is an iterative procedure, with each iteration comprising two steps, denoted E (expectation) and M (maximization). The M step is the computational procedure for the complete dataset, in which the functions of the missing data are replaced by their conditional (posterior) expectations given the current estimates of the parameters and the data. These conditional expectations are evaluated in the E step immediately preceding the M step.

In the context of responses to multiple-choice items, the E step involves modelling of the hypothetical responses to replace the missing ones. First, the hypothetical (complete-data) responses attain values of zero or unity, and so their distributions are completely determined by conditional probabilities. Thus, the task at hand is to establish the conditional probabilities of correct (hypothetical) response for all the missing responses.

In the model implied by the current operational procedures, each 'omit' corresponds to not knowing, and so the test taker would have chosen a response option (out of K choices) at random. However, the test-taker may be able to eliminate some of the options, in which case the (hypothetical) chance of a correct response would be higher than $1/K$. In principle, the estimate of the guessing parameter c_i in (1), can be used as the posterior probability of correct response for item i. However, these probabilities are usually estimated with considerable sampling variation; the data contain only limited information about the guessing parameters.

The first missing response in a sequence of 'not reached' is likely to be an 'omit' (the item was attended), say, with probability r_i, the second response in

such a sequence is somewhat less likely to an 'omit'; its chance is less or equal to $\min(r_i, r_{i+1})$. Assuming that ocurrences of omission are positively correlated this probability is greater than $r_i r_{i+1}$. Responses further down the sequence are 'omits' with diminishing probabilities.

Not reaching is assumed to be non-informative. That is, the chance of the correct (hypothetical) response to an item that has not been attended is the same as the chance among test-takers with the same proficiency who did attend the item. The proficiency estimates would be available from the previous M step, or from the starting solution.

Evaluation of the various joint probabilities of correct response is enabled by the conditional independence structure of the responses.

The EM algorithm described in this section is computationally very demanding because its M step is in itself a complex iterative procedure (often slow to converge). Even though later iterations of the EM algorithm may comprise a single iteration of the M step, a large number of iterations may be required, especially when a lot of responses are missing.

The next section presents an EM algorithm based on a different complete dataset. Its principal advantages are that its M step is considerably less complex and the E step is more flexible.

Knowledge states as the complete data

In the EM algorithm described in the previous section the dichotomous format is 'forced' on the complete data. It seems counter to intuition that an omitted response (which implies 'DK') is converted into 'C' with a certain probability, instead of incorporating the information that the test-taker does not know the correct response. This line of thought naturally leads to considering a different EM algorithm, in which the knowledge states ('K' and 'DK') are the complete data.

Now the M step involves the two-parameter item-response model

$$\{p_i(\theta_t) \ = \} \quad \mathrm{P}(R_{ti} = \mathrm{K}) \ = \ \mathrm{logit}^{-1}(a_i + b_i\theta_t), \tag{2}$$

obtained from (1) by setting $c_i = 0$. This model is much easier to fit, with item parameters a_i and b_i and proficiencies θ_t random or fixed, than that in (1).

The E step involves estimation of the knowledge state given the response. Omitted and incorrect responses correspond to 'DK' with certainty. When the response is correct or not reached there is a lot of uncertainty about the knowledge state.

Models for correct responses

In the E step of the EM algorithm outlined above the conditional probabilities of knowing given the response (other than incorrect) have to be evaluated. For that a probabilistic model relating response categories to knowledge states must be posited. In brief, the conditional probabilities of correct responses, as functions of the item, test-taker's proficiency, response category, and possibly some other features, have to be specified.

Suppose test-taker t responded correctly to item i. If he/she has low proficiency, it is likely to have been a good guess; a test-taker with high proficiency most likely knew the correct answer. In general,

$$\mathrm{P}(R_{ti} = \mathrm{K} \mid \mathrm{C}) \;=\; \frac{p_i(\theta_t)}{c_i + (1 - c_i)p_i(\theta_t)}.$$

Models for 'not-reached' responses

It is very difficult to make an educated guess about the proficiency of a test-taker who abandons the test. The information contained in the responses up to the item where he/she abandoned the test has to be relied on. The model underlying the current operational procedures is that the abandoning is conditionally non-informative given the test-taker's proficiency. This enables a relatively simple scheme for imputing for 'not-reached' responses.

The assumption of non-informative abandoning cannot be tested and is not supported by any educational measurement theory. Inability to respond to an item, disinterest in the tested subject matter, and the like, are possible causes

of abandoning the test. Most of such causes are negatively associated with proficiency.

It is important to make a distinction between 'not-reaching' as used in the operational classification of responses, and abandoning a test (not attending any of a sequence of items that includes the last item). The first 'not-reached' response may be an 'omit'; an 'omit' cannot be followed by a sequence of 'not-reached'. A test-taker may be less likely to abandon the test immediately after an item with knowledge state 'K' than after an item with 'DK'.

These considerations lead to a mixture model for 'not-reached' responses. For each 'not-reached' response a conditional probability is specified as a function of the response's position in the sequence of 'not-reached', the last presented response, test-taker's proficiency, distance from the end of the test, and the like. Information about the parameters of such a model is very diffuse, or has to rely on educated guess. For instance, it would be natural to posit that the conditional probability of knowledge state 'K' for an 'abandoned' item is somewhat lower than the probability given by (2). An 'omit' is more likely to be followed by abandoning or by another 'omit'. Insight about the relative frequencies of these sequences of responses (and the underlying conditional probabilities) can be gained only from large-scale administrations, or possibly, by pooling information across such administrations (e.g., of the same test), on surveys of test-takers (taking into account the low reliability of their responses to questions about their testing behaviour), and possibly incorporating untestable assumptions.

Estimation of subpopulation means without IRT

Applications of the item-response theory with incomplete data discussed above highlight the importance of the appropriate choice of the *complete* dataset. Estimation of the subpopulation means of proficiencies can also be looked upon as an application of the EM algorithm. For the M-step, we consider estimation of the subpopulation mean with the values of the proficiencies given, while the E-step comprises estimation of the proficiency values. In practice, the sampling

variation of the estimated proficiency scores is represented by set of five plausible values.

The plausible values are an intermediate step in the estimation of subpopulation means. It would be desirable to have models and associated computational routines that relate summaries of the item responses of a subpopulation to the subpopulation mean of the ability, on a scale not necessarily identical with the scale implied by the item response theory. Such an approach would bypass the computationally tedious estimation of the ability of each test-taker and accumulation of random errors in these estimates.

Summary

An approach for incorporation of information about missing responses is outlined. It relies on a model relating knowledge categories (know, or does not know) to the response categories (correct, incorrect, omitted, not reached, multiple). A computational algorithm is described which requires no new technology to be developed.

References

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B 39, 1–38.

Wainer, H. (1985). Estimating the number of examinees who did not reach the last item of a section. ETS Technical Report 85–60. Educational Testing Service, Princeton, NJ.

Table 1: Classification for responses and knowledge. Possible combinations.

	Response classes			
	Response		Missing response	
Knowledge classes	C	N	O	A
K	Yes	No	No	Yes
DK	Yes	Yes	Yes	Yes

Table 2: Crosstabulation of response types for items 5 and 6 in Reading block C.

Item 6 / Item 5	Respond	Omit	Not reached	Total
Respond	911	576	11	1498
Omit	16	30	0	46
Not-reached	0	0	27	27
Total	927	606	38	1571

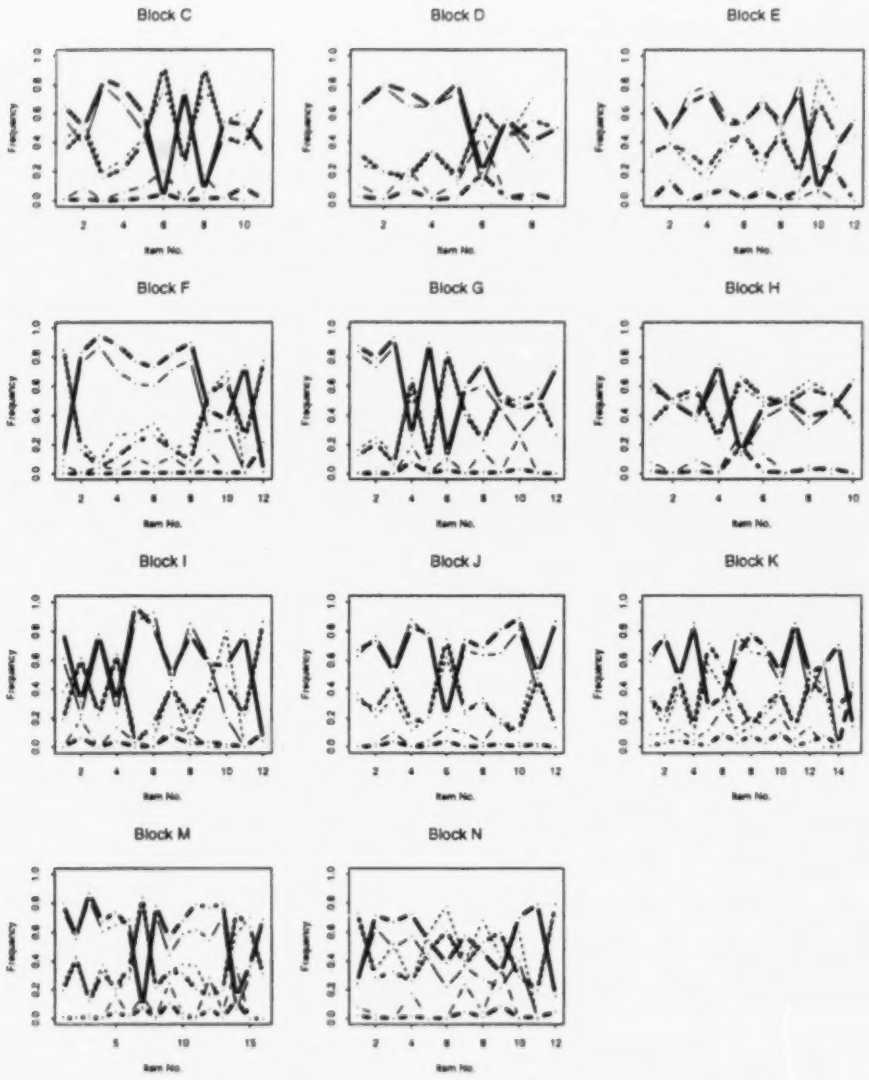

Figure 1: Proportions of correct, incorrect and omitted for test-takers who reached and those who failed to failed to reach the last item.
Each plot represents a block. The horizontal axis of each plot is the item order number, the vertical axis the observed proportion of correct (solid line), incorrect (dotted line), and omit (dashed line). The thick lines are for those who reached the last item, the thin lines for those who did not.

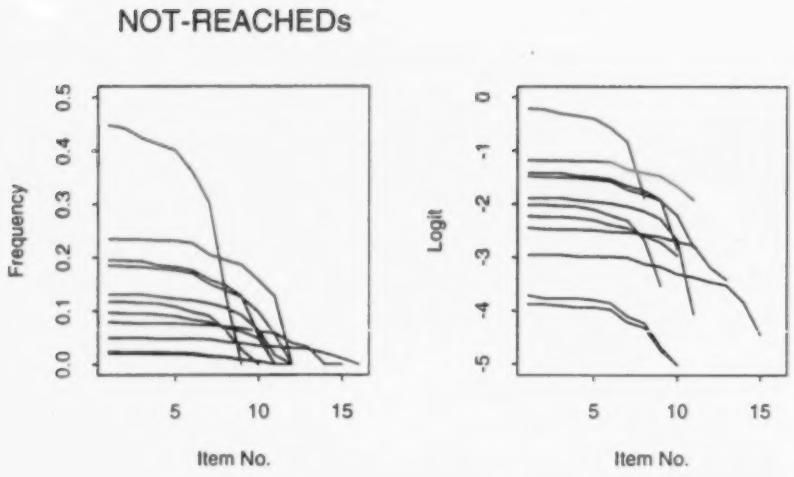

Figure 2: Proportions of not reached.
Each line represents a block. For each item (horizontal axis) the proportion of test-takers who reached the item but did not reach the last item, is plotted. In the left-hand-side plot the proportion scale, and in the right-hand side plot the logit scale is used.